

Enhanced Fisher Linear Discriminant Models for Face Recognition

Chengjun Liu and Harry Wechsler
Department of Computer Science, George Mason University,
4400 University Drive, Fairfax, VA 22030-4444, USA
{cliu, wechsler}@cs.gmu.edu

Abstract

We introduce in this paper two Enhanced Fisher Linear Discriminant (FLD) Models (EFM) in order to improve the generalization ability of the standard FLD based classifiers such as Fisherfaces. Similar to Fisherfaces, both EFM models apply first Principal Component Analysis (PCA) for dimensionality reduction before proceeding with FLD type of analysis. EFM-1 implements the dimensionality reduction with the goal to balance between the need that the selected eigenvalues account for most of the spectral energy of the raw data and the requirement that the eigenvalues of the within-class scatter matrix in the reduced PCA subspace are not too small. EFM-2 implements the dimensionality reduction as Fisherfaces do. It proceeds with the whitening of the within-class scatter matrix in the reduced PCA subspace and then chooses a small set of features (corresponding to the eigenvectors of the within-class scatter matrix) so that the smaller trailing eigenvalues are not included in further computation of the between-class scatter matrix. Experimental data using a large set of faces — 1,107 images drawn from 369 subjects and including duplicates acquired at a later time under different illumination — from the FERET database shows that the EFM models outperform the standard FLD based methods.

1. Introduction

A successful face recognition methodology depends heavily on the particular choice of the features used by the (pattern) classifier [2], [5]. Feature selection in pattern recognition involves the derivation of salient features from the raw input data in order to reduce the amount of data used for classification and simultaneously provide enhanced discriminatory power. Two popular techniques for selecting a subset of features are Principal Component Analysis (PCA) and the Fisher Linear Discriminant (FLD). PCA is a standard decorrelation technique and following its application one derives an orthogonal projection basis which directly

leads to dimensionality reduction, and possibly to feature selection. Methods such as FLD, possibly following methods such as PCA and operating then in a compressed subspace, seek for discriminatory features by taking into account within- and between-class scatter as the relevant information for pattern classification. While PCA maximizes for all the scatter as appropriate for signal representation, FLD differentiates between the within- and between-class scatter as appropriate for pattern classification.

PCA was first applied to represent human faces by Kirby and Sirovich [4] and to recognize faces by Turk and Pentland [7]. The recognition method, known as eigenfaces, defines a feature space, or “face space”, which drastically reduces the dimensionality of the original space, and face detection and identification are carried out in the reduced space.

FLD has also been applied to solve face recognition problems. By applying first PCA for dimensionality reduction and then FLD for discriminant analysis Belhumire, Hespanha, and Kriegman [1] developed an approach called Fisherfaces for face recognition. Using a similar approach, Swets and Weng [6] have pointed out that the eigenfaces derived using PCA are only the most expressive features (MEF). The MEF are unrelated to actual face recognition, and in order to derive the most discriminating features (MDF), one needs a subsequent FLD projection.

Methods that combine PCA and standard FLD, however, lack in their generalization ability as they overfit to the training data. This paper explains the reasons why overfitting takes place, introduces two Enhanced FLD Models (EFM) to overcome problems associated with overfitting, and presents experimental data whose enhanced performance supports the EFM models.

2. The Standard FLD Based Methods and Overfitting

The standard FLD based methods such as Fisherfaces apply first PCA for dimensionality reduction and then discriminant analysis. Relevant questions concerning PCA are

usually related to the range of Principal Components (PCs) used and how it affects performance. Regarding discriminant analysis one has to understand the reasons for overfitting and how to avoid it. The answers to those two questions are closely related. One can actually show that using more PCs may lead to decreased performance (for recognition). The explanation for this behavior is that the trailing eigenvalues correspond to high-frequency components and usually encode noise. As a result, when these trailing but small valued eigenvalues are used to define the reduced PCA subspace, the FLD procedure has to fit for noise as well and as a consequence overfitting takes place.

The FLD procedure involves the simultaneous diagonalization of the two within- and between-class scatter matrices and it is stepwise equivalent to two operations as pointed out by Fukunaga [3]: first whitening the within-class scatter matrix, and second applying PCA on the between-class scatter matrix using the transformed data. The purpose of the whitening step is to normalize (to unity) the within-class scatter matrix for uniform gain control. The second operation maximizes then the between-class scatter to separate different classes as much as possible. The robustness of the FLD procedure depends on whether or not the within-class scatter captures reliable variations for a specific class. Note that when PCA precedes FLD for dimensionality reduction and more PCs are used, the trailing eigenvalues of the within-class scatter matrix tend to capture noise and their values are fairly small. As during whitening the eigenvalues of the within-class scatter matrix appear in the denominator, the small (trailing) eigenvalues cause the whitening step to fit for misleading variations and thus generalize poorly when exposed to new data. The Enhanced FLD Models (EFM) presented in the next section address those problems and display increased generalization ability.

3. Enhanced FLD Models (EFM)

Two Enhanced FLD Models (EFM-1 and EFM-2) are introduced to improve on the generalization ability of the standard FLD based methods such as Fisherfaces. Both EFM-1 and EFM-2 apply first PCA for dimensionality reduction before proceeding with FLD type of analysis. EFM-1 implements the dimensionality reduction step with the goal to balance between the need that the selected eigenvalues (corresponding to the PCs for the original image space) account for most of the spectral energy of the raw data and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA subspace) are not too small. EFM-2 implements dimensionality reduction as Fisherfaces do. It proceeds with the whitening of the within-class scatter matrix in the reduced PCA subspace and then chooses a small set of features (corresponding to the eigenvectors of the within-class scatter matrix) so that the smaller

trailing eigenvalues are not included in further computation of the between-class scatter matrix.

3.1. Dimensionality Reduction (PCA)

PCA generates a set of orthonormal basis vectors, known as principal components (PCs), that maximize the scatter of all the projected samples. Let $X = [X_1, X_2, \dots, X_n]$ be the sample set of the original images. After normalizing the images to unity norm and subtracting the grand mean a new image set $Y = [Y_1, Y_2, \dots, Y_n]$ is derived. Each Y_i represents a normalized image with dimensionality N , $Y_i = (y_{i1}, y_{i2}, \dots, y_{iN})^t$, ($i = 1, 2, \dots, n$). The covariance matrix of the normalized image set is defined as

$$\Sigma_Y = \frac{1}{n} \sum_{i=1}^n Y_i Y_i^t = \frac{1}{n} Y Y^t \quad (1)$$

and the eigenvector and eigenvalue matrices Φ , Λ are computed as

$$\Sigma_Y \Phi = \Phi \Lambda \quad (2)$$

Note that $Y Y^t$ is an $N \times N$ matrix while $Y^t Y$ is an $n \times n$ matrix. If the sample size n is much smaller than the dimensionality N , then the following method saves some computation [7]

$$(Y^t Y) \Psi = \Psi \Lambda_1 \quad (3)$$

$$\mathfrak{S} = Y \Psi \quad (4)$$

where $\Lambda_1 = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$, and $\mathfrak{S} = [\Phi_1, \Phi_2, \dots, \Phi_n]$. If one assumes that the eigenvalues are sorted in decreasing order, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, then the first m leading eigenvectors define matrix P

$$P = [\Phi_1, \Phi_2, \dots, \Phi_m] \quad (5)$$

The new feature set Z with lower dimensionality m ($m \ll N$) is then computed as

$$Z = P^t Y \quad (6)$$

3.2. EFM-1

For EFM-1 model, our purpose is to choose m , the number of PCs in Eq. 5, such that proper balance is preserved between the need that the selected eigenvalues (corresponding to the PCs for the original image space) account for most of the spectral energy of the raw data and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA subspace) are not too small. The eigenvalue spectrum of PCA is derived by Eq. 3, and the relative magnitude of the eigenvalues for the face data used in our experiments (see Sect. 4) is shown in Fig. 1. The index for the eigenvalues ranges from 1 to $m = L$, where L stands

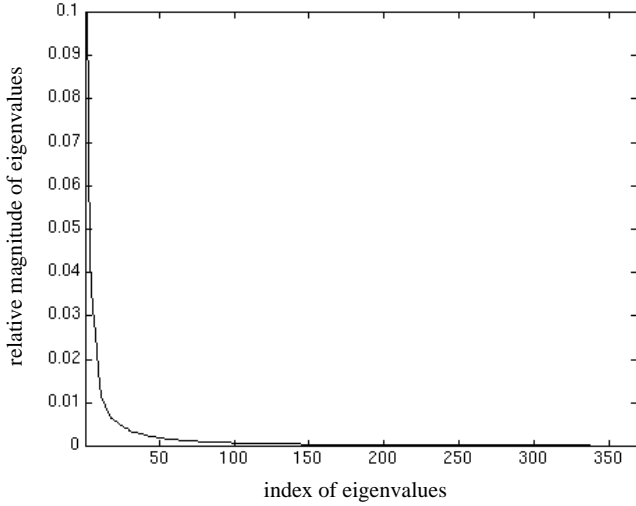


Figure 1. The relative magnitude $(\lambda_i / \sum_{k=1}^m \lambda_k)$ **of eigenvalues**

for the number of (face) classes considered in our experiments. One can see from Fig. 1 that the first 50 eigenvalues capture most of the energy and that the eigenvalues whose index is greater than 100 are fairly small and most likely capture noise.

To calculate the eigenvalue spectrum of the within-class scatter matrix in the reduced PCA subspace one needs to decompose the FLD procedure into two operations: whitening and diagonalization [3]. In particular, let $\omega_1, \omega_2, \dots, \omega_L$ and N_1, N_2, \dots, N_L denote the classes and the number of images within each class, respectively. Let M_1, M_2, \dots, M_L and M be the means of the classes and the grand mean in the reduced PCA subspace $\text{span}[\Phi_1, \Phi_2, \dots, \Phi_m]$. We then have

$$M_k = \frac{1}{N_k} \sum_{j=1}^{N_k} Z_j^{(k)}, \quad k = 1, 2, \dots, L \quad (7)$$

$$M = \sum_{k=1}^L P(\omega_k) M_k \quad (8)$$

where $Z_j^{(k)}$, $j = 1, 2, \dots, N_k$, represents the sample images from class ω_k , and $P(\omega_k)$ is the prior probability of ω_k . The within- and between-class scatter matrices Σ_w and Σ_b in the reduced PCA subspace are estimated as

$$\Sigma_w = \sum_{k=1}^L P(\omega_k) \left\{ \frac{1}{N_k} \sum_{j=1}^{N_k} \left(Z_j^{(k)} - M_k \right) \left(Z_j^{(k)} - M_k \right)^t \right\} \quad (9)$$

$$\Sigma_b = \sum_{k=1}^L P(\omega_k) (M_k - M)(M_k - M)^t \quad (10)$$

The standard FLD procedure derives a projection matrix A that maximizes the ratio $|A^t \Sigma_b A| / |A^t \Sigma_w A|$. The ratio is maximized when A consists of the eigenvectors of $\Sigma_w^{-1} \Sigma_b$ corresponding to the leading eigenvalues. The step-wise FLD procedure derives the eigenvalues and eigenvectors of $\Sigma_w^{-1} \Sigma_b$ as the result of the simultaneous diagonalization of Σ_w and Σ_b . First whiten the within-class scatter matrix

$$\Sigma_w \Xi = \Xi \Gamma \quad \text{and} \quad \Xi^t \Xi = I \quad (11)$$

$$\Gamma^{-1/2} \Xi^t \Sigma_w \Xi \Gamma^{-1/2} = I \quad (12)$$

The eigenvalue spectrum of the within-class scatter matrix in the reduced PCA subspace can be derived by Eq. 11, and different spectra are obtained corresponding to different number of PCs that are utilized. Fig. 2 displays the relative magnitudes of the eigenvalues corresponding to the within-class scatter matrix. Now one has to simultaneously optimize the behavior of the trailing eigenvalues in the reduced PCA space (Fig. 2) with the energy criteria for the original image space (Fig. 1). Note that different choices on the cutoff PC index for the original image space yield different within-class spectra as shown in Fig. 2. As one can see from Fig. 1 and 2, a suitable choice is to set $m = 50$, since this choice not only accounts for most of the spectral energy of the raw data (Fig. 1) but also meets the requirement that the eigenvalues of the within-class scatter matrix (in the reduced 50 dimensional PCA subspace) are not too small (Fig. 2).

After the number of PCs is set, EFM-1 model computes the between-class scatter matrix as

$$\Gamma^{-1/2} \Xi^t \Sigma_b \Xi \Gamma^{-1/2} = K_b \quad (13)$$

Diagonalize the new between-class scatter matrix K_b

$$K_b \Theta = \Theta \Delta \quad \text{and} \quad \Delta^t \Delta = I \quad (14)$$

The overall transformation matrix (after PCA, see Eq. 6) for EFM-1 model is now defined as

$$T = \Xi \Gamma^{-1/2} \Theta \quad (15)$$

3.3. EFM-2

The EFM-2 model uses the same number of PCs as utilized by Fisherfaces [1] and MDF method [6] for dimensionality reduction, namely, $L \leq m \leq n - L$, where n is the number of training samples and L the number of classes.

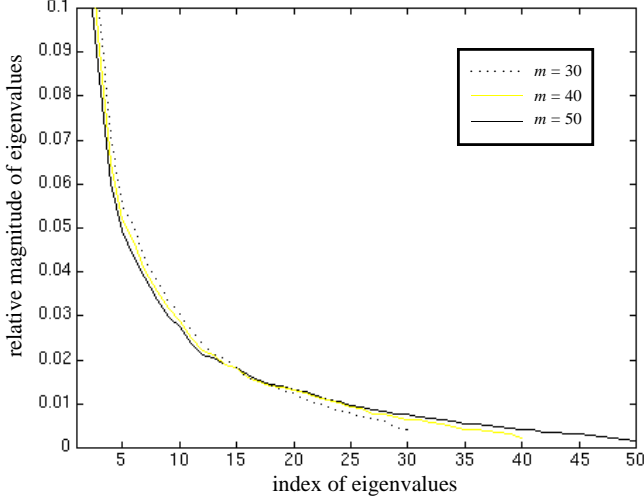


Figure 2. The relative magnitude of the eigenvalues spectra derived using the FLD whitening step with different number of PCs ($m = 30$, $m = 40$, and $m = 50$) for dimensionality reduction

The FLD proceeds then by choosing a small set of features (corresponding to the eigenvectors of the within-class scatter matrix — see Eq. 11) after the whitening procedure so that the smaller trailing eigenvalues are not included.

Let s ($s < m$) be the number of features chosen in the whitening subspace according to the eigenvalue spectrum of Σ_w (see Fig. 3, a suitable choice is to set $s = 100$), and the new feature matrix Ξ^* is derived

$$\Xi^* = [\xi_1, \xi_2, \dots, \xi_s] \quad (16)$$

where $\xi_1, \xi_2, \dots, \xi_s$ are the eigenvectors of Σ_w corresponding to the leading eigenvalues $\gamma_1, \gamma_2, \dots, \gamma_s$. The new diagonal matrix is

$$\Gamma^* = \text{diag} \{ \gamma_1, \gamma_2, \dots, \gamma_s \} \quad (17)$$

The new whitening transformation matrix in the reduced s dimensional whitening subspace is

$$Q = (\Xi^*) (\Gamma^*)^{-1/2} \quad (18)$$

The between-class scatter matrix Σ_b in this subspace is transformed to K_b^*

$$K_b^* = Q^t \Sigma_b Q \quad (19)$$

Note that the new between-class scatter matrix K_b^* is $s \times s$ now instead of $m \times m$. Diagonalize K_b^*

$$K_b^* \Theta^* = \Theta^* \Delta^* \quad \text{and} \quad \Theta^{*t} \Theta^* = I \quad (20)$$

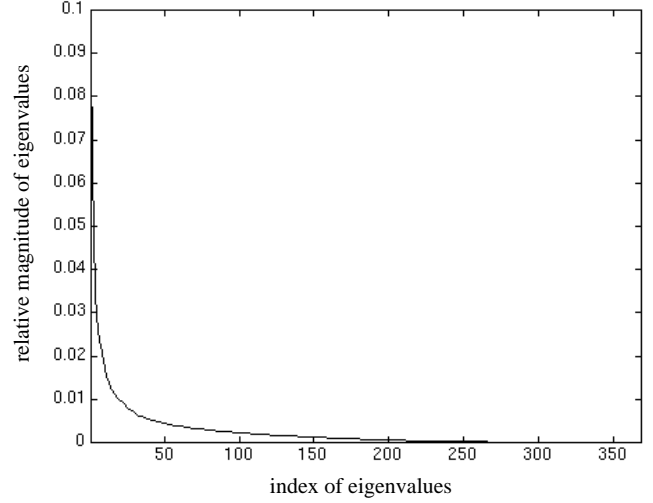


Figure 3. The relative magnitude ($\lambda_i / \sum_{k=1}^m \lambda_k$) of the eigenvalues derived by the whitening FLD step

The overall transformation matrix (after PCA, see Eq. 6) for EFM-2 model is

$$T^* = \Xi^* \Gamma^{*-1/2} \Theta^* \quad (21)$$

4. Experimental Results and Future Work

The experimental data, consisting of 1,107 facial images corresponding to 369 subjects, comes from the FERET database. 600 out of the 1,107 images correspond to 200 subjects with each subject having three images — two of them are the first and the second shot, while the third shot is taken under low illumination. For the remaining 169 subjects there are also three images for each subject, but two out of the three images are duplicates taken at a different time. Two images of each subject are used for training with the remaining one used for testing. The images are cropped to the size of 64×96 , once the eye coordinates are manually detected.

Fig. 4 displays the recognition performance for the Fisherfaces. It confirms that after using a certain number of features (around 70) the performance peaks. The graphs were obtained when m is optimally set to L because $L \leq m \leq n - L$ and $n = 2L$. The top 1 recognition rate records the accuracy rate for the top response being correct, while the top 3 recognition rate records the accuracy rate for the correct response being included among the first three ranked choices.

Fig. 5 shows the comparative performance for Fisherfaces and our new EFM-1 and EFM-2 models. The EFM

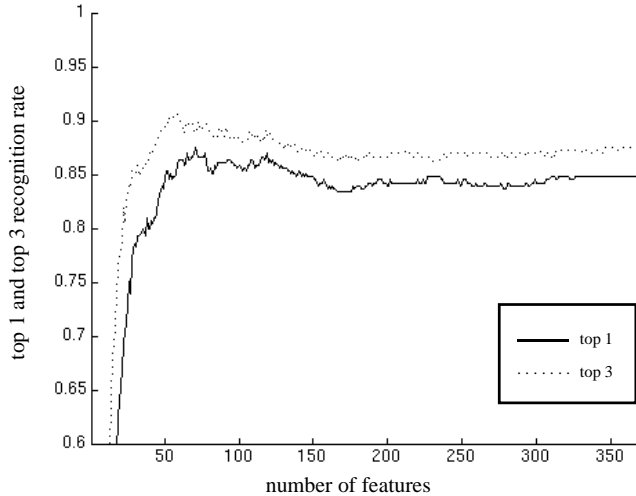


Figure 4. The recognition performance for Fisherfaces with 369 PCs ($m = 369$)

models increase the top 1 recognition rate by 10% to 15% compared to the Fisherfaces. Fig. 6 plots the top 3 recognition rate, and again our EFM models improve the performance by 10% to 15%.

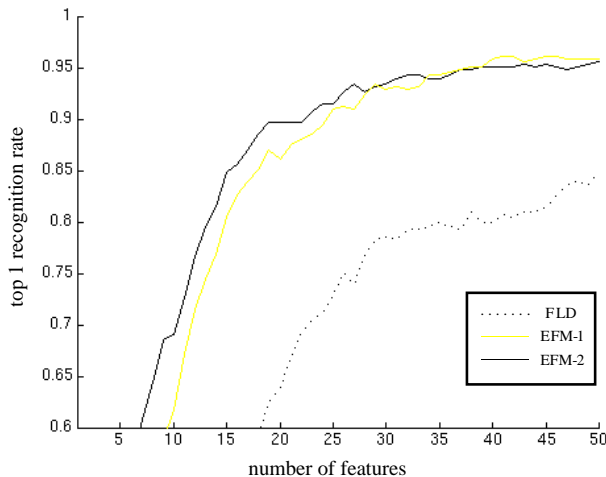


Figure 5. The comparative performance (top 1) for Fisherfaces, EFM-1 and EFM-2 models

We plan to expand the EFM approach so that it seeks the best subset of PCs rather than the first ones corresponding to the leading eigenvalues. While the first PCs are probably useful as they encode for global image characteristics, in analogy to the characteristics displayed by the human visual system, there is reason to believe that one should at least attenuate the very first PCs. As PCA encodes 2nd order

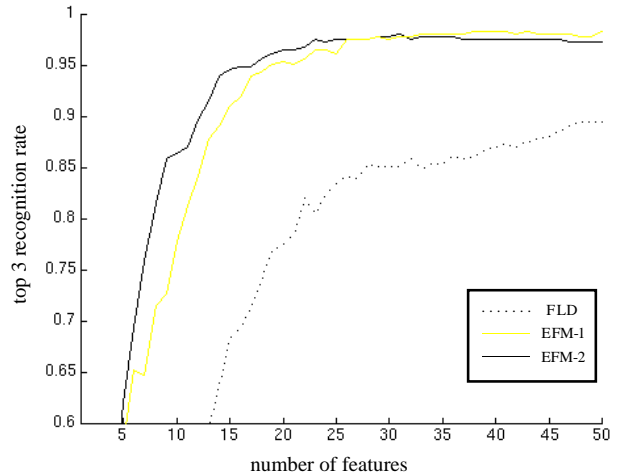


Figure 6. The comparative performance (top 3) for Fisherfaces, EFM-1 and EFM-2 models

statistics and those are equivalent to the power spectrum, one possible choice would be to weight the PCs with a filter whose characteristics are similar to the Contrast Sensitivity Function (CSF). Another possibility is to actually find the best subset of PCs using a greedy search type of algorithm.

Acknowledgments: This work was partially supported by the DoD Counterdrug Technology Development Program, with the U.S. Army Research Laboratory as Technical Agent, under contract DAAL01-97-K-0118.

References

- [1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
- [2] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proc. IEEE*, 83(5):705–740, 1995.
- [3] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1991.
- [4] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.
- [5] A. Samal and P. Iyengar. Automatic recognition and analysis of human faces and facial expression: A survey. *Pattern Recognition*, 25(1):65–77, 1992.
- [6] D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. on PAMI*, 18(8):831–836, 1996.
- [7] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 13(1):71–86, 1991.